Explainable AI

Introduction to Artificial Intelligence and Explainability

IARIA 2021, 80118

Summary

Artificial Intelligence (AI) utilises algorithms to make decisions or support human decision making by analysing huge data sets, finding patterns, and proposing courses of action, and they do this at scales beyond human capability [7][8]. AI has the ability to transform every industry sector by imitating and augmenting human intelligence and removing inconsistencies in human decision making [8][9][17]. While AI has a long way to go before it can reason abstractly about real-world situations and achieve the human level of reasoning and social interaction, it is responsible for the highly lucrative online consumer industry and for bringing science-fiction into reality with driverless cars [8][17].

Explainability is a major barrier to acceptance and utilisation of AI [1][20]. "The current generation of AI systems offer tremendous benefits, but their effectiveness is constrained the machine's inability to explain its decisions and actions" [6]. This is most apparent in more conservative industry sectors, such as banking and finance, health and security, where the penetration of AI is nominal. Explainable AI will be essential if industry leaders, professional specialists and other AI stakeholders are to understand, appropriately trust, and effectively manage this incoming generation of artificially intelligent partners.

This Lecture introduces Artificial Intelligence and Machine Learning. We look at some of the areas where AI has been successfully applied, and analyse why AI is not being utilized more broadly. Following that we investigate different methods for making AI and ML explainable, and approaches for interpreting ML models, measuring their performance, and generating greater user acceptance amongst the AI stakeholders.

The video for this tutorial is available from: https://youtu.be/4nfc0kPkAtU